Requested Patent: JP11120183A

Title: METHOD AND DEVICE FOR EXTRACTING KEYWORD ;

Abstracted Patent: JP11120183 ;

Publication Date: 1999-04-30 ;

Inventor(s): HARA MASAMI;; KITANI TSUYOSHI ;

Applicant(s): NTT DATA CORP ;

Application Number: JP19970276093 19971008 ;

Priority Number(s): ;

IPC Classification: G06F17/30 ;

Equivalents: ;

ABSTRACT:

PROBLEM TO BE SOLVED: To provide a keyword extracting device which automatically extracts a keyword from text (computerized document data) which is divided into items also with high accuracy.SOLUTION: A significance processing part 15 decides item significance based on learning words that exist in each item from a learning text group which is preliminarily classified in each category and its different item appearance frequency and holds a set of item significance about each category. When new text is inputted, word significance of each new word is decided based on a set of item significance which is held about a category to which the new text belongs and the different item appearance frequency of the new word which is extracted from the new text and transfers it to a keyword deciding part 17. The part 17 specifies the prescribed number of new words whose word significance is relatively high as a keyword and outputs it to an output device.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－120183

(43)公開日　平成11年(1999)４月30日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３１０Ａ
			15/403	３４０Ａ

審査請求　未請求　請求項の数８　ＯＬ　（全７頁）

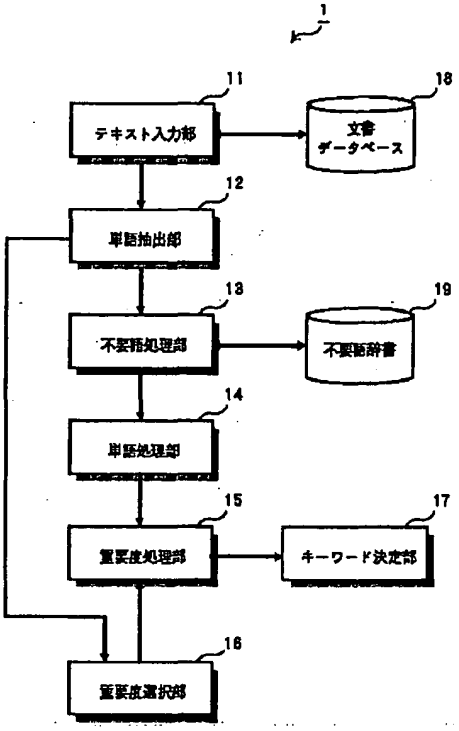
(21)出願番号　特願平9－276093

(22)出願日　平成９年(1997)10月８日

(71)出願人　000102728
株式会社エヌ・ティ・ティ・データ
東京都江東区豊洲三丁目３番３号
(72)発明者　原　正巳
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内
(72)発明者　木谷　強
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内
(74)代理人　弁理士　鈴木　正剛

(54)【発明の名称】　キーワード抽出方法及び装置

(57)【要約】
【課題】　項目分けされたテキスト（電子化された文書データ）からキーワードを自動的且つ高精度に抽出できるキーワード抽出装置を提供する。
【解決手段】　重要度処理部１５において、予めカテゴリ毎に分類された学習用テキスト群から各項目に存する学習単語及びその項目別出現頻度に基づく項目重要度を決定し、個々のカテゴリについての項目重要度の組を保持しておく。新規テキストが入力された場合は、その新規テキストが属するカテゴリについて保持されている項目重要度の組とその新規テキストから抽出した新規単語の項目別出現頻度とに基づいて各新規単語の単語重要度を決定してキーワード決定部１７に渡す。キーワード決定部１７は、単語重要度が相対的に高い所定数個の新規単語をキーワードとして特定して図示しない出力装置に出力する。



【特許請求の範囲】
【請求項1】　それぞれ項目分けされてカテゴリ毎に分類された学習用テキスト群から各項目に存する学習単語及びその項目別出現頻度に基づく項目重要度を決定するとともに、個々のカテゴリについての項目重要度の組を保持しておき、
項目分けされてカテゴリ毎に分類された新規テキストからその特徴を表すキーワードを抽出する際に、当該新規テキストが属するカテゴリについて保持されている前記項目重要度の組と前記新規テキストから抽出した新規単語の項目別出現頻度とに基づいて各新規単語の単語重要度を決定し、この単語重要度が相対的に高い新規単語を前記キーワードとして特定することを特徴とするキーワード抽出方法。
【請求項2】　前記項目重要度の決定は、各カテゴリの第1項目に存する第1単語の出現頻度を当該カテゴリ内のすべての項目に存する第1単語の出現頻度で除算して頻度割合を算出し、予め第1単語に付与された単語重要度と前記頻度割合との乗算値を第1項目に存するすべての単語について算出し、その算出結果を加算することによって行うことを特徴とする請求項1記載のキーワード抽出方法。
【請求項3】　前記新規単語の単語重要度の決定は、前記新規テキストが属するカテゴリ内で、第2項目に存する複数の新規単語の中から特定した第2単語の出現頻度と当該カテゴリにおける第2項目の項目重要度とを乗算することにより行うことを特徴とする請求項1記載のキーワード抽出方法。
【請求項4】　予め定めた不要単語と同一の新規単語を削除した後に前記単語重要度を決定することを特徴とする請求項1記載のキーワード抽出方法。
【請求項5】　前記単語重要度に閾値を定めて複数個の新規単語が前記キーワードとして特定されるようにし、これらのキーワードによって当該新規テキストの概要を把握できるようにしたことを特徴とする請求項1記載のキーワード抽出方法。
【請求項6】　項目分けされてカテゴリ毎に分類された複数のテキストからそれぞれ単語を抽出し、抽出した個々の単語に検索重みを表す単語重要度を付与する手段と、キーワード抽出対象となる新規テキストから抽出された複数の新規単語のうち前記キーワードとすべき1または複数の新規単語を特定するキーワード特定手段とを備えたキーワード抽出装置において、
前記キーワード特定手段は、
学習に用いる前記テキストから各項目に存する学習単語及びその項目別出現頻度に基づく項目重要度を決定するとともに、個々のカテゴリについての項目重要度の組を保持しておき、前記新規テキストが属するカテゴリについて保持されている前記項目重要度の組と前記新規単語の項目別出現頻度とに基づいて各新規単語の単語重要度を決定し、この単語重要度が相対的に高い新規単語を前記キーワードとして特定するように構成されていることを特徴とするキーワード抽出装置。
【請求項7】　不要単語を登録した不要語辞書と、
前記新規単語に前記不要語辞書中の不要単語と同一のものが含まれているときに当該新規単語を削除する不要単語削除手段とをさらに備え、前記不要単語が削除された新規単語に対して前記単語重要度を決定することを特徴とする請求項6記載のキーワード抽出装置。
【請求項8】　前記キーワード特定手段は、前記単語重要度が高い所定数個の新規単語を前記キーワードとして特定することを特徴とする請求項6記載のキーワード抽出装置。
【発明の詳細な説明】
【0001】
【発明の属する技術分野】本発明は、テキスト（電子文書データ、以下同じ）から、そのテキストの特徴を表すキーワードを自動的に抽出し、利用者が容易に必要な情報を検索できるようにするためのキーワード抽出技術に関する。
【0002】
【従来の技術】従来、キーワードの抽出は、人間がテキストを熟読して内容を熟知したうえで手作業で行う必要があった。しかし、近年のように多種の情報の電子化が進むにつれ、計算機によりキーワードを自動的且つ効果的に抽出する方法が検討されている。このような方法の一つに、計算機上で文の意味を解釈する意味理解技術を用いて、文中における各語の重要性を決定する方法が知られている。一方、上記意味理解技術を用いた方法とは異なり、テキストの表層情報を利用することによって文中における各語の重要性を決定する研究も行われている。例えば、複合語の接続関係や出現頻度、語長等を考慮して、キーワードの重要性を決定している方法（亀田、「キーワード抽出装置」：特開平8－95982号公報）や、単語分割を行った後で、出現順序が連続する単語の組を複合語としてとらえ、それらの語において出現頻度や出現件数の多い語をキーワードとする手法（神林、清水他、「インターネット情報検索に適したキーワード抽出」、情処学会自然言語処理研究会〔1997〕）等が提案されている。
【0003】また、従来のキーワード抽出手法では、テキスト全体を一括して処理対象とするための検討がなされている。しかし、テキスト内の段落や文章の重要性は全て異なり、キーワードの抽出には適切ではない段落や文章も存在することになる。したがって、キーワードを抽出する際には、重要な箇所の的確な選択が必要である。このことに関連して、同一カテゴリのテキストでは、重要な箇所が類似しているという性質を利用して、キーワードを抽出する手法がいくつか提案されている。このような手法は、例えば、「キーワード自動抽出にお

ける分野特性の利用：木本、電子情報通信学会全国大会〔1989〕」や、「出現度数と分野情報を利用したキーワード抽出法の検討：原、情報処理学会〔1991〕」等に詳述されている。これらの提案では、テキスト内の重要段落の位置が、テキストの属するカテゴリに応じて変化するものとしており、当該カテゴリに応じて決定した段落の重要度を、段落内における単語の重要度に対して反映させている。

【0004】

【発明が解決しようとする課題】しかし、上述の提案では、重要な段落の決定は全て人間が行っている。そのため、キーワード抽出に際しては、決定者の経験や判断に依存するところが大きい。また、決定者によっては、特定の段落のみが選択されてキーワードが抽出されるために、人為処理に起因して該抽出時におけるキーワードの漏れや不足等が生じるおそれがあった。

【0005】そこで本発明の課題は、テキストからキーワードを自動的に抽出し、利用者が、容易に必要な情報を検索して活用できるようにする、改良されたキーワード抽出方法を提供することにある。また、本発明の他の課題は、上記キーワード抽出方法の実施に好適なキーワード抽出装置を提供することにある。

【0006】

【課題を解決するための手段】上記課題を解決する本発明のキーワード抽出方法は、それぞれ項目分けされてカテゴリ毎に分類された学習用テキスト（学習に用いるテキスト、以下同じ）群から各項目に存する学習単語及びその項目別出現頻度に基づく項目重要度を決定するとともに、個々のカテゴリについての項目重要度の組を保持しておく。そして、項目分けされてカテゴリ毎に分類された新規テキストからその特徴を表すキーワードを抽出する際に、当該新規テキストが属するカテゴリについて保持されている前記項目重要度の組と前記新規テキストから抽出した新規単語の項目別出現頻度とに基づいて各新規単語の単語重要度を決定し、この単語重要度が相対的に高い新規単語を前記キーワードとして特定することを特徴とする。

【0007】前記項目重要度の決定は、例えば、各カテゴリの第1項目に存する第1単語の出現頻度を当該カテゴリ内のすべての項目に存する第1単語の出現頻度で除算して頻度割合を算出し、予め第1単語に付与された単語重要度と前記頻度割合との乗算値を第1項目に存するすべての単語について算出し、その算出結果を加算することによって行う。

【0008】また、前記新規単語の単語重要度の決定は、例えば、前記新規テキストが属するカテゴリ内で、第2項目に存する複数の新規単語の中から特定した第2単語の出現頻度と当該カテゴリにおける第2項目の項目重要度とを乗算することにより行う。

【0009】なお、前記単語重要度の決定は、予め定めた不要単語と同一の新規単語を削除した後に行い、また、前記単語重要度に閾値を定めて複数個の新規単語が前記キーワードとして特定されるようにして、これらのキーワードによって当該新規テキストの概要を把握できるようにすることが好ましい。

【0010】上記他の課題を解決する本発明のキーワード抽出装置は、項目分けされてカテゴリ毎に分類された複数のテキストからそれぞれ単語を抽出し、抽出した個々の単語に検索重みを表す単語重要度を付与する手段と、キーワード抽出対象となる新規テキストから抽出された複数の新規単語のうち前記キーワードとすべき1または複数の新規単語を特定するキーワード特定手段とを備えたキーワード抽出装置において、前記キーワード特定手段が、学習に用いる前記テキストから各項目に存する学習単語及びその項目別出現頻度に基づく項目重要度を決定するとともに、個々のカテゴリについての項目重要度の組を保持しておき、前記新規テキストが属するカテゴリについて保持されている前記項目重要度の組と前記新規単語の項目別出現頻度とに基づいて各新規単語の単語重要度を決定し、この単語重要度が相対的に高い新規単語（1または複数個）を前記キーワードとして特定するように構成されていることを特徴とする。

【0011】好ましい実施の形態として、不要単語を登録した不要語辞書と、前記新規単語に前記不要語辞書中の不要単語と同一のものが含まれているときに当該新規単語を削除する不要単語削除手段とをさらに備え、前記不要単語が削除された新規単語に対して前記単語重要度を決定するようにする。

【0012】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を詳細に説明する。図1は、本発明を適用したキーワード抽出装置の一実施形態を示す機能ブロック図である。このキーワード抽出装置1はコンピュータ装置によって実現されるもので、そのコンピュータ装置の内部あるいは外部記憶装置に構築される文書データベース18及び不要語辞書19と、上記コンピュータ装置が所定のプログラムを読み込んで実行することにより形成される、テキスト入力部11、単語抽出部12、不要語処理部13、単語処理部14、重要度処理部15、重要度選択部16、キーワード決定部17、とを備えて構成される。

【0013】なお、上記プログラムは、通常、コンピュータ装置の内部あるいは外部記憶装置に格納され、随時読み取られて実行されるようになっているが、コンピュータ装置に所用の機能を付与できる形態のものであれば、それがどのように記録されているかは特に問題とはならない。例えば、コンピュータ装置と分離可能な可搬性記録媒体、例えばCD-ROM（コンパクトディスク型ROM）、FD（フレキシブルディスク）、MD（ミニディスク）等に格納され、使用時に上記内部または外

部記憶装置にインストールされて随時実行に供されるものであってもよい。

【0014】文書データベース18には、検索対象となるテキスト（学習テキスト／新規テキスト）が読み出し自在に蓄積されている。これらのテキストは、予め、項分け記載された定型フォーマットにしたがって複数のカテゴリ毎に分類されている。この場合の定型フォーマットは、“項目1”、“項目2”、…“項目N”のN項目から構成されるものである。

【0015】不要語辞書18には、利用者がキーワードの抽出に際して「不要」と判断した不要単語群が登録されている。なお、この不要語辞書18中の不要単語群も上述の定型フォーマットに基づいて各項目及び各カテゴリ毎に分類されて登録されているものである。

【0016】テキスト入力部11は、文書データベース18中の学習テキスト、及び追加される新規テキストの入力を受け付け、受け付けたテキストを単語抽出部12に渡す。単語抽出部12は、テキストに対して所定の形態素解析を施し、名詞句に相当する単語の抽出を行う。学習用テキストからは学習単語を抽出し、新規テキストからは新規単語を抽出する。新規単語については、これを不要語処理部13に入力する。

【0017】不要語処理部13は、単語抽出部12で抽出された複数の新規単語と不要語辞書18中に登録された単語との照合を行うもので、不要語辞書18に同一の新規単語が存在する場合はその新規単語を削除する。同一の新規単語が不要語辞書19中に存在しない場合は、複数の新規単語をそのまま単語処理部14に渡す。

【0018】単語処理部14は、入力された各単語に対して、それぞれ該当テキストに対する検索重みを表す単語重要度を付与するものである。この単語重要度は、検索処理等で広く利用されている「TF・IDF法（G.Salton他：“Introduction toModern Information Retrieval”,McGraw-Hill）」や、「χ^2分布」を利用した方法（長尾他：「日本語文献における重要度の自動抽出」情報処理、Vol.17,No.2,情報処理学会）等の手法に基づいて付与することができる。単語重要度が付与された各単語は重要度処理部15に入力される。

【0019】重要度処理部15は、単語重要度が付与された各単語から、テキストの定型フォーマットにおける項目毎の重みを表す項目重要度を算出し、算出した項目重要度の組を図示しないデータ記憶手段にカテゴリ毎に保持しておく。また、新規テキストが属するカテゴリが特定されたときに、そのカテゴリをキーとしてデータ記憶手段から該当する項目重要度の組を選択し、選択した項目重要度の組と新規テキストから抽出した新規単語の項目別出現頻度とに基づいて各新規単語の単語重要度を算出する。算出結果はキーワード決定部16に入力される。

【0020】キーワード決定部17は、単語重要度が相対的に高い1または複数個の新規単語をキーワードとして決定し、図示しない出力装置に出力するものである。

【0021】次に、上記キーワード抽出装置1の全体の動作を具体的に説明する。まず、文書データベース18に蓄積されているテキストのうち、学習用テキスト群から項目重要度を決定する。この学習用テキスト群には、予めカテゴリC1、C2…、CLが付与されているものとする。

【0022】図2は、学習用テキスト群に含まれる個々の学習単語に付与される単語重要度の決定手順を示す概略図である。カテゴリC_k（$1 \leq k \leq L$）が付与されている学習用テキスト群は、テキスト入力部11に入力される。単語抽出部12は、この学習用テキスト群を形態素解析し、全学習用テキストから単語切り出しを行う（ステップS1）。切り出された単語群から名詞句に相当する単語を抽出して学習単語（W_1～W_M）とし、これらを単語処理部14に渡す。単語処理部14は、学習単語（W_1～W_M）の各々に対し、上述の「TF・IDF法」等によって、正規化された単語重要度（S_1～S_M）を付与する（ステップS2）。これらの単語重要度（S_1～S_M）に基づいて項目重要度を算出する過程を示したのが図3である。この処理は、重要度処理部15においてなされる。すなわち、重要度処理部15は、ステップS2において付与された単語重要度（S_1～S_M）に基づき、以下のようにして項目重要度を算出する。まず、学習単語W_1～W_Mの単語重要度を各々S_1～S_Mとした場合のカテゴリC_kにおける項目j（$1 \leq j \leq N$）の重要度$D_{k,j}$を以下に示す式（1）で定義する。

【0023】

【数1】

$$D_{kj} = \sum_{W_i \in \text{項目}j} s_i * \frac{freq^j(W_i)}{\sum_{j=1}^{N} freq^j(W_i)} \quad \cdots(1)$$

【0024】ここで、式（1）の分子（$freq^j(W_i)$）は、項目j内における学習単語W_iの出現頻度、また、式（1）の分母（$\Sigma freq^j(W_i)$）は、全項目における単語W_iの出現頻度である。また、$freq^j(W_i) / \Sigma freq^j(W_i)$は、項目jにおける単語$W_i$の出現頻度とテキスト全体における単語$W_i$の出現頻度との割合（以下、出現頻度割合）を表す。この出現頻度割合に対して、学習単語W_iの単語重要度S_iを乗じたものが、学習単語W_iに関する項目jの重要度である。この重要度を、項目jに存する全ての単語について加算し、これをカテゴリC_kにおける項目jの重要度$D_{k,j}$とする。以上の処理によって決定したカテゴリC_kにおける項目重要度の組をD_k（$=(D_{k,1}, D_{k,2}, \cdots, D_{k,N})$）とする（ステップS3）。重要度処理部15において算出された重要度の組は、キーワード決定部17によりキーワードを決定する際に用いられる。

【0025】次に、図4を参照して、キーワード未抽出

テキスト、すなわち新規テキストからキーワードを抽出する手順を説明する。なお新規テキストにおいても、予め上述同様のカテゴリが付与されているものとする。

【0026】上記学習用テキストの場合と同様に、テキスト入力部11に入力された新規テキストから単語抽出部12で新規単語が抽出される。抽出された新規単語群に不要語が含まれている場合は、不要語処理部13において削除される（ステップS4）。不要語が削除された新規単語群は、重要度選択部16に入力され、新規テキストの属するカテゴリC_t（但し、$1 \leqq t \leqq L$）の項目重要度の組$D_t=(D_{t,1}, D_{t,2}, \cdots, D_{t,N})$が選択される。そして、重要度処理部15において、新規単語群に対してそれぞれ単語重要度を付与する（ステップS5）。このとき新規単語W_1（但し、lは、1，2，・・・）に付与される単語重要度A（W_1）は、以下の式（2）で定義される。

【0027】

【数2】

$$A(W_l)=\sum_j D_{t,j} * freq^j(W_l) \quad \cdots(2)$$

【0028】ここで、$D_{t,j}$は、カテゴリC_tにおける項目jの重要度であり、$freq^j$（W_1）は、項目j内における新規単語W_1の出現頻度である。以上のようにして新規単語に付与された重要度A（W_1）（$1 \leqq l \leqq T$：Tは新規テキストに出現する全単語数）に基づいて、キーワード決定部17では、単語重要度A（W_1）を例えば降順に並べて新規単語（の種類）を分類・整列させる。そして、より上位に位置する所定数個の新規単語を、抽出すべきキーワードとして決定する（ステップS6）。

【0029】なお、本実施形態のキーワード抽出装置1では、文書データベース18に追加される新規テキストが逐次テキスト入力部11に入力されるものとしているが、文書データベース18中への新規テキストの追加を自動検出する新規テキスト検出手段を文書データベース18とテキスト入力部11との間に介在させるようにしてもよい。このようにすれば、上記キーワード抽出処理を自動化することが可能になる。

【0030】このように、本実施形態のキーワード抽出装置1では、新規テキストに対して項目別の重要度を加味し、さらに、カテゴリ毎に異なる項目の重要性を考慮して新規単語に重要度を付与し、当該重要度に基づいてキーワードを抽出するようにしたので、キーワード抽出の精度を従来よりも格段に高めることができる。

【0031】抽出されたキーワードは、例えば、検索システムにおける検索キーとして利用可能となる。また、抽出されたキーワードは、単語重要度に基づいて決定されているので、当該新規テキストの概要を表現していることになる。したがって、利用者は、対象となるテキスト全文を読むことなく、当該キーワードを読むだけで当該新規テキストに関する情報の概要を把握することができるようになる。

【0032】さらに、キーワードの抽出を自動化することにより、従来のように、人間が判定していた重要な段落の決定等が省略可能となるとともに、人為処理に起因していたキーワードの漏れや不足等が回避できるようになる。

【0033】

【発明の効果】以上の説明から明らかなように、本発明によれば、新規テキストから、その特徴をより良く表すキーワードが自動的に抽出されるので、利用者は、抽出されたキーワードを利用することにより、容易に必要な情報を検索して活用できるようになる。また、キーワードは、新規テキストにおける項目の重要度に基づいて抽出されることから、高精度のキーワード抽出が可能となる。さらに、本発明を検索システム等に適用させた場合には、処理効率及びその実用性が格段に向上するシステムの提供が可能になる。

【図面の簡単な説明】

【図1】本発明の一実施形態に係るキーワード抽出装置における機能ブロック図。

【図2】学習用テキスト群における単語重要度の決定手順を示す概略図。

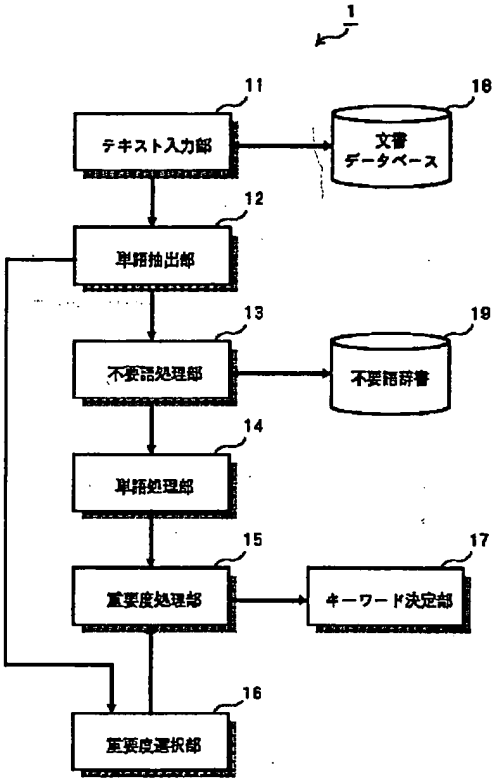
【図3】項目重要度の決定手順を示す概略図。

【図4】新規テキストからのキーワード抽出手順を示す概略図。

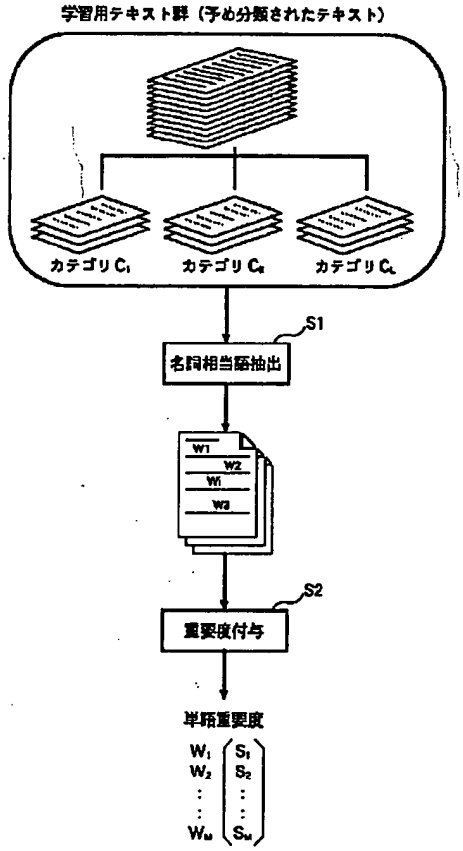
【符号の説明】

1　キーワード抽出装置
11　テキスト入力部
12　単語抽出部
13　不要語処理部
14　単語処理部
15　重要度処理部
16　重要度選択部
17　キーワード決定部
18　文書データベース
19　不要語辞書

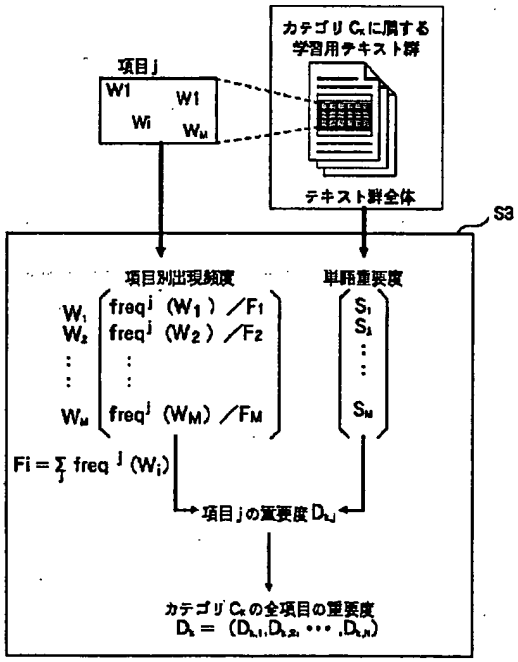
【図1】


1
11
テキスト入力部
18
文書
データベース
12
単語抽出部
13
不要語処理部
19
不要語辞書
14
単語処理部
15
重要度処理部
17
キーワード決定部
16
重要度選択部

【図2】


学習用テキスト群（予め分類されたテキスト）
カテゴリ C1
カテゴリ Ck
カテゴリ CK
S1
名詞相当語抽出
W1
W2
Wi
W3
S2
重要度付与
単語重要度
W1 S1
W2 S2
Wi Si
WM SM

【図3】


項目j
W1 W1
Wi WM
カテゴリ Ck に関する
学習用テキスト群
テキスト群全体
S3
項目別出現頻度
単語重要度
W1 freqj (W1) /F1 S1
W2 freqj (W2) /F2 S2
WM freqj (WM) /FM SM
Fi = Σj freqj (Wi)
項目jの重要度 Dkj
カテゴリ Ck の全項目の重要度
Dk = (Dk1, Dk2, ・・・, DkN)